Apport et défis des Big Data en cancérologie

Pierre Saintigny [1,2], Jean-Philippe Foy [1,3], Anthony Ferrari [4], Philippe Cassier [2], Alain Viari [4,5], Alain Puisieux [1]

1. Université Lyon, université Claude-Bernard Lyon 1, centre Léon-Bérard, centre de recherche en cancérologie de Lyon, CNRS 5286, Inserm 1052, 69008 Lyon, France
2. Centre Léon-Bérard, département de médecine et département de recherche translationnelle et de l'innovation, 28, rue Laënnec, 28, rue Laennec, 69373 Lyon cedex 08, France
3. Université Pierre-et-Marie-Curie, Paris 6, hôpital Pitié-Salpêtrière, service de chirurgie maxillo-faciale, groupe hospitalier Pitié-Salpêtrière, 47-83, boulevard de l'Hôpital, 75013 Paris cedex 13, France
4. Plateforme de bio-informatique–Gilles-Thomas, synergie Lyon Cancer, 69008 Lyon, France
5. INRIA Grenoble-Rhône-Alpes, 655, avenue de l'Europe, 38330 Montbonnot-Saint- Martin, France

Correspondance :
Pierre Saintigny, cancer research center of Lyon, CNRS 5286, Inserm U1052, 28, prom Léa-et-Napoléon-Bullukian, 69008 Lyon, France.
pierre.saintigny@lyon.unicancer.fr

Mots clés
Big Data
Cancérologie
Classification moléculaire
Hétérogénéité
Médecine de précision

Résumé

Depuis le premier séquençage du génome humain en 2001, le développement de nouvelles technologies à haut débit, ainsi que la diminution considérable du coût du séquençage ont permis des avancées importantes en oncologie. La caractérisation moléculaire des cancers a notamment permis d'identifier des anomalies oncogéniques clés au cours du processus tumoral, permettant le développement de stratégies thérapeutiques personnalisées. Cependant, la quantité d'information considérable ainsi générée a créé de nouveaux défis à relever comme le stockage, le traitement ou encore l'exploitation de l'information. Dans cet article, nous décrivons l'apport et les défis représentés par les Big Data en cancérologie.

Keywords
Big Data
Oncology
Molecular classification
Heterogeneity
Precision medicine

Summary

Contribution and challenges of Big Data in oncology

Since the first draft of the human genome sequence published in 2001, the cost of sequencing has dramatically decreased. The development of new technologies such as next generation sequencing led to a comprehensive characterization of a large number of tumors of various types as well as to significant advances in precision medicine. Despite the valuable information this technological revolution has allowed to produce, the vast amount of data generated resulted in the emergence of new challenges for the biomedical community, such as data storage, processing and mining. Here, we describe the contribution and challenges of Big Data in oncology.

Introduction

Chaque année, à travers le monde, 5,3 millions d'hommes et 4,7 millions de femmes développent une tumeur maligne et 6,2 millions de personnes décèdent d'un cancer. Ainsi, le cancer constitue l'un des enjeux humains, socioéconomiques, médicaux et scientifiques majeurs en ce début de XXI^{e} siècle. En France, les cancers représentent aujourd'hui la première cause de mortalité et les conséquences sociétales de ce fléau ont conduit à l'élaboration de trois Plans Cancer, dont les objectifs sont de garantir une prise en charge d'excellence pour les patients, de développer la recherche et les actions de prévention et promouvoir les applications des dernières avancées diagnostiques et thérapeutiques vers la clinique. Face à la fréquence, à la complexité et à l'hétérogénéité des pathologies tumorales, l'enjeu est le développement d'une médecine de précision permettant à la fois d'améliorer la survie globale tout en limitant la morbidité associée aux traitements. La définition des pathologies cancéreuses comme des « pathologies génétiques », c'est-à-dire qui se développent sur la base de l'accumulation progressive d'anomalies affectant l'ADN des cellules pré-tumorales et tumorales, a conduit, au cours des dix dernières années, au développement de grands programmes internationaux de caractérisation moléculaire des cancers, avec l'objectif de dresser un catalogue exhaustif des anomalies génétiques et épigénétiques impliquées dans les différents types de cancers. Cette ambition a fait entrer de plain-pied la biologie des cancers dans le domaine ubiquitaire des Big Data.

Généralités sur la biologie des cancers

Même si le cancer a été reconnu dès la fin du XIX^{e} siècle comme une maladie résultant d'un dérèglement du fonctionnement des chromosomes, l'identification, il y a une trentaine d'années, des premiers gènes impliqués dans la transformation maligne a permis d'élaborer une théorie unifiée concernant les mécanismes moléculaires mis en jeu dans la genèse et le développement d'une tumeur : la progression tumorale correspond à un processus dynamique qui tend à sélectionner un clone cellulaire présentant une ou plusieurs altérations génétiques favorisant sa survie et son expansion. Le développement d'une tumeur repose donc sur un processus complexe qui n'est pas lié à l'altération isolée d'un gène, mais à l'apparition progressive d'altérations moléculaires liées entre elles ou indépendantes touchant un grand nombre de gènes et protéines, conférant directement un avantage sélectif pour la cellule ou indirectement, par la dérégulation d'autres gènes et protéines. La compréhension de ce processus et la caractérisation d'une tumeur donnée nécessitent donc d'identifier les événements génétiques et épigénétiques altérés dans une tumeur donnée, de décrypter les réseaux de signalisation intra- et intercellulaires et de comprendre leur conséquence biologique tant au niveau cellulaire que tissulaire. En ce sens, le déchiffrage du génome humain et les nouveaux outils technologiques d'analyse de l'ADN et de l'ARN tumoraux ont ouvert des perspectives majeures pour mieux caractériser les tumeurs et évaluer avec précision leur pronostic clinique. L'identification de gènes clefs, altérés de façon récurrente dans les cancers, a également permis au cours de ces dernières années tout à la fois de caractériser les tumeurs avec une meilleure précision, et donc d'améliorer la qualité du diagnostic et du pronostic, et de proposer de nouvelles cibles thérapeutiques, à l'origine des « thérapeutiques ciblées », avec l'ambition d'évoluer vers une médecine de précision.

Les Big Data en cancérologie

Parmi les 4 « V » définissant les Big Data et correspondant à la variété, la véracité, le volume et la vélocité, les deux plus importants sont sans aucun doute, le volume et la variété des données produites. Ces données sont apparues dans le domaine biomédical à la fin des années 1990 avec l'émergence des puces à ADN permettant l'analyse du produit de l'expression de l'ensemble des gènes (entre 25 000 et 50 000 dans le cas du génome humain) d'un échantillon biologique (par exemple un échantillon d'une tumeur de patient opéré) et la publication de la séquence complète du génome humain (près de 3 milliards de nucléotides) en 2001. Depuis, les progrès technologiques, considérables dans le domaine du séquençage de l'ADN et de l'ARN, ont conduit à la production d'une quantité de données importantes (de l'ordre de plusieurs dizaines de Péta octets). Citons à titre d'exemple les dépôts de données transcriptomiques Gene Expression Omnibus et ArrayExpress [1,2], qui s'enrichissent en permanence à l'occasion de la publication des travaux des chercheurs et ce à la demande des éditeurs des revues scientifiques. Actuellement, chacune de ces bases représente plus de deux millions d'échantillons, pour la plupart issues d'expériences sur puces, mais également, et de plus en plus, de séquençage.

Mais c'est dans le domaine du séquençage ADN (DNASeq) et ARN (RNASeq) que les progrès sont les plus considérables, avec la montée en puissance des techniques *next generation sequencing* (NGS). Dans le domaine du cancer, les deux plus importantes initiatives sont celles du The Cancer Genome Atlas (TCGA) et de l'International Cancer Genome Consortium (ICGC). Le TCGA est essentiellement nord-américain et résulte d'une collaboration entre le National Cancer Institute (NCI) et le National Human Genome Research Institute (NHGRI). Il porte actuellement sur une trentaine de types de cancers rassemblant les échantillons de plus de 11 000 patients pour un volume total de données d'environ 2,5 Péta octets. L'ICGC est quant à lui une initiative internationale rassemblant près d'une vingtaine de pays (dont les US, une partie des données du TCGA sont incluses dans l'ICGC) sur une vingtaine de types de cancer et près de 15 000 patients. Dans les deux cas, l'objectif est de caractériser de manière exhaustive et systématique un grand nombre d'échantillons de cancers (~500 échantillons par

type), ces échantillons étant pour la majorité prélevés sur des pièces opératoires de patients au diagnostic [3,4]. À travers les financements de l'Institut national du cancer (INCa) et de l'Inserm, la France participe à l'ICGC sur huit pathologies : le sein (sous-type amplifié sur HER2), le foie, la prostate, le sarcome d'Ewing, le rétinoblastome, les carcinosarcomes gynécologiques et le léiomyosarcome (http://www.e-cancer.fr/Professionnels-de-la-recherche/Innovations/Les-progres-de-la-genomique/ICGC-France). Les données produites par ces projets sont constituées généralement par des séquences de génomes complets (*whole genome sequencing* [WGS]) avec, et ceci est une particularité du séquençage dans le domaine du cancer, un échantillon tumoral (à une couverture moyenne de 50X) et un échantillon normal issu du même individu (généralement sanguin, à une couverture de 30X) afin d'identifier les altérations somatiques, par soustraction des altérations constitutionnelles. Dans certains cas, le séquençage est limité aux régions codantes (on parle alors de *whole exome sequencing* [WXS]). Elles sont, dans tous les cas, associées au séquençage de l'ARN tumoral (RNASeq) et complétées par des analyses plus spécifiques (Methylome par puces ou séquençage bisulfite, MiRNome, etc.). Les informations apportées par le séquençage portent sur l'identification d'altérations ponctuelles (*single nucleotide variants* [SNV]), de nombre de copies de gènes (*copy number variations* [CNV]) ou de grands réarrangements intra- et interchromosomiques (*structural variants* [SV]). Il est important de noter qu'en plus des altérations portant sur les gènes (oncogènes ou suppresseurs de tumeurs), les variations somatiques non géniques (accessibles en WGS) fournissent également d'importantes informations sur les processus ayant conduit à la formation ou au développement des cellules cancéreuses. Ainsi le contexte nucléotidique d'un SNV porte une « signature » caractéristique du processus carcinogène initial, par exemple par les UV (dans les mélanomes) ou un agent exogène (tabac) [5]. Près d'une trentaine de signatures caractéristiques ont ainsi été identifiées à ce jour et sont référencées dans la base Cosmic [5]. La base Ensembl (*variant effect predictor* [VEP]) permet également de prédire les conséquences d'un variant nucléotidique sur le gène touché ou encore la séquence protéique attendue. Cette notion de signature (initialement définie sur les SNV) a récemment été étendue au cas des variants structuraux dans une étude, à laquelle la France a participé, portant sur 560 génomes complets de cancers du sein [6]. Un autre exemple récent illustratif de l'utilisation des données WGS au-delà des régions géniques concerne l'analyse (à partir des données WGS) du processus d'amplification d'ERBB2 dans les cancers du sein HER2-positif [7].

Il est important de noter que l'accès à toutes ces données est rendu public au travers de portails Web dédiés pour tout ce qui concerne les données somatiques et d'expression, non identifiantes et de faible volumétrie (http://www.cbioportal.org/ ; https://gdc-portal.nci.nih.gov/ ; https://dcc.icgc.org/). En revanche, l'accès aux données constitutionnelles, et a fortiori, aux génomes complets, fait l'objet d'autorisations spécifiques délivrées par les comités d'accès aux données (Data Access Committee) et réclame du côté des utilisateurs potentiels, d'importantes ressources de calcul et de stockage, situées bien au-delà de ce dont le biologiste ou clinicien dispose couramment. La question de la facilité d'accès à ces données, notamment au travers du cloud, aux utilisateurs finaux est encore ouverte et constitue un enjeu majeur des prochaines années.

Ces programmes ont permis une évolution exponentielle du nombre d'études en cancérologie ayant su tirer profit de ces données (*figure 1*) [8]. Si la révolution scientifique représentée par l'accès public à ces données « à haut débit », se conçoit facilement, l'exploitation et l'utilisation pratique de ces données par le chercheur peuvent parfois paraître plus difficiles. Ainsi, l'expression « chercher une aiguille dans une botte de foin » peut prendre tout son sens quand il s'agit d'extraire l'information concernant un gène, un transcrit ou une protéine en particulier, au sein de la complexité de l'ensemble du génome, du transcriptome ou du protéome. La création d'interfaces Web comme cBioPortal [9,10] et « The Cancer Genome Atlas Clinical Explorer » [11], ou encore la mise au point de packages utilisables en langage R comme TCGA2STAT [12], sont autant d'outils qui permettent de limiter l'écart existant parfois entre l'exhaustivité des Big Data et les questions des chercheurs. Ainsi, de plus en plus d'outils voient le jour pour faire face à l'évolution exponentielle de ces données (*figure 1*). Ils représentent probablement les clés nécessaires pour l'utilisation exhaustive de l'ensemble des données dans le domaine de la recherche. Cependant, de nouveaux outils devront être développés pour répondre aux nouveaux challenges à venir en cancérologie comme l'interprétation des données croissantes de méta-génomique pour la caractérisation du microbiote dans les différents types de cancers.

Enfin, en parallèle de ces vastes programmes de collection de données cliniques et moléculaires, les données générées par séquençage ciblé à visée diagnostique peuvent également représenter un défi technique et humain : technique, par le stockage et la gestion de données de plus en plus importantes dans des laboratoires de taille plus modeste, et humain par la nécessité de recruter un personnel compétent pour l'analyse bio-informatique de celles-ci. Ainsi, le principal défi en « routine clinique » sera également, outre le fait de générer ces données, de les exploiter pour personnaliser la stratégie thérapeutique du patient.

Classification moléculaire des cancers

Ces grands programmes internationaux ont permis de proposer des classifications moléculaires de nombreuses tumeurs humaines, principalement basées sur les données transcriptomiques, mais intégrant également des données de mutations somatiques affectant des oncogènes ou des gènes suppresseurs de

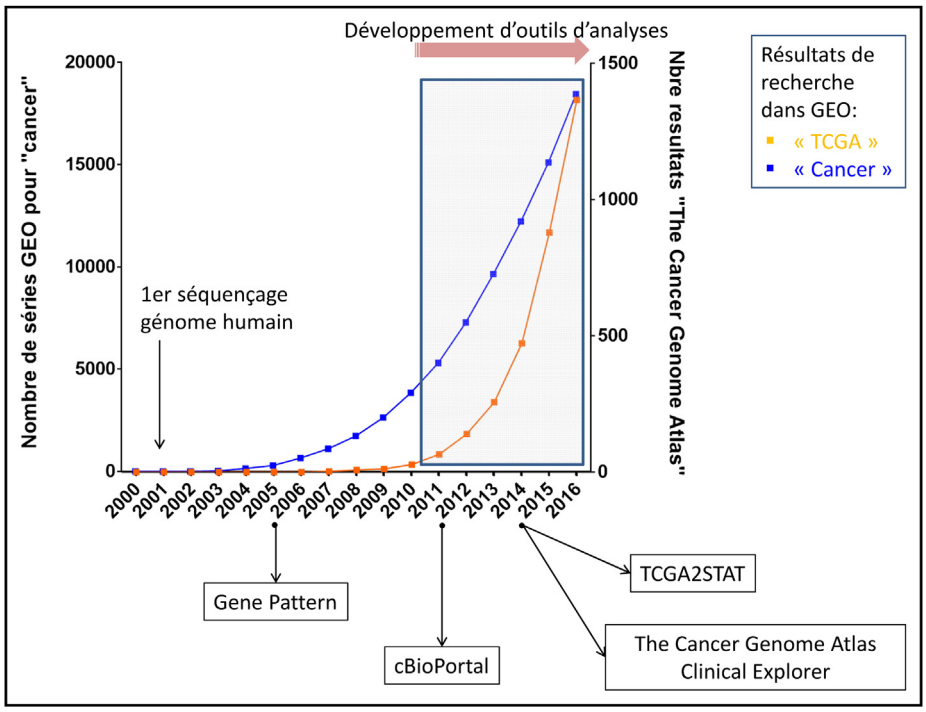

FIGURE 1

Évolution des données à haut débit de 2000 à 2016. Nombre d'études par année, obtenues par la recherche des mots « cancer » dans Gene Expression Omnibus (GEO) et « The Cancer Genome Atlas » (TCGA) dans PubMed

tumeurs, les altérations du nombre de copies de gènes, les altérations de la méthylation de l'ADN entre autres. D'une manière générale, ces classifications moléculaires coïncident partiellement avec les informations anatomocliniques utilisées en pratique clinique quotidienne. Leur incorporation dans la prise en charge des patients reste cependant à démontrer pour la plupart. Elles doivent pour cela montrer un intérêt pronostic supérieur aux paramètres simples actuellement utilisés en clinique, ou montrer qu'ils sont prédictifs de la sensibilité ou de la résistance à telle ou telle approche thérapeutique. Les classifications moléculaires des cancers participent néanmoins à une meilleure compréhension de la biologie des tumeurs et fournissent une grille de lecture aux chercheurs pour valider la pertinence clinique des phénomènes étudiés au laboratoire.

À titre d'exemple, la classification appelée « PAM50 » des cancers du sein permet de distinguer cinq sous-types moléculaires (luminal A ou B, basal, HER2-enriched et normal-*like*) sur la base du produit de l'expression de 50 gènes [13]. Plus récemment une classification en 10 groupes a été proposée intégrant les données de variation du nombre de copies de gènes [14]. Une classification du cancer colorectal propose de distinguer 4 sous-types moléculaires [*consensus molecular subtypes* (CMSs)] : CMS1 (*microsatellite instability immune*), CMS2 (*canonical*), CMS3 (*metabolic*), et CMS4 (*mesenchymal*) [15]. Pour le cancer de la prostate, 7 sous-types ont été proposés [16], quatre étant caractérisés par des gènes de fusion impliquant la famille des facteurs de transcription ETS (ERG, ETV1, ETV4 et FLI1), et trois autres étant définis par des mutations des gènes *SPOP*, *FOXA1* et *IDH1*. Il faut noter que 25 % des cancers de la prostate ne peuvent être classés dans l'un de ces 7 sous-types. Les cancers bronchiques non à petites cellules qui représentent 85 % de l'ensemble des cancers pulmonaires étaient classés exclusivement sur la base de l'histologie jusqu'au début des années 2000 en adénocarcinome, carcinome épidermoïde et carcinome à grandes cellules. La caractérisation moléculaire de ces tumeurs a permis de reclasser les carcinomes à grandes cellules en adénocarcinomes ou carcinomes épidermoïdes, d'identifier au sein des adénocarcinomes pulmonaires plusieurs « drivers » oncogéniques plus fréquents chez les patients non-fumeurs (EGFR, ALK, ROS1, HER2, BRAF...), c'est-à-dire des gènes dont les altérations génomiques, exclusives les unes des autres, activent de façon constitutive des voies de signalisation pour lesquelles les cellules tumorales ont une addiction [17–19]. Cette classification est à la base du traitement de première ligne des patients ayant une maladie métastatique. Les carcinomes épidermoïdes pulmonaires et ceux des voies aérodigestives supérieures (VADS) sont quant à eux classés en 4 groupes (pulmonaires : *classical*, *primitive*, *basal* and *secretory ;* VADS :

atypical, basal, classical, mesenchymal), classification n'ayant pour l'instant aucun impact dans la prise en charge des patients [20].

L'accumulation de ces données dans de multiples types tumoraux différents permet l'analyse transversale de divers processus biologiques dans une approche « pan-cancer » [21]. Dans le contexte de l'émergence rapide de l'immunothérapie des cancers, un exemple marquant est le développement d'approches pan-cancer pour apprécier le contexte immunitaire intratumoral [22,23].

Hétérogénéité intratumorale temporelle et spatiale et plasticité cellulaire

Les données générées par les grands programmes du TCGA et de l'ICGC ont mis récemment en exergue la complexité remarquable de ces lésions, et ont fortement ravivé l'intérêt pour l'hétérogénéité intratumorale. La diversité des cellules cancéreuses au sein d'une même tumeur est connue depuis plusieurs décennies par les pathologistes. Dans une large mesure, cette hétérogénéité peut être attribuée à la plasticité et aux capacités d'adaptation des cellules cancéreuses qui font notamment intervenir des processus épigénétiques. Le concept d'évolution « Darwinienne » appliqué au niveau cellulaire a permis de proposer que les tumeurs malignes proviennent d'une cellule unique (origine monoclonale des cancers) par l'acquisition progressive de mutations oncogéniques conférant un avantage sélectif de prolifération ou de survie cellulaire menant à des vagues successives d'expansion clonale [23,24]. Cette évolution repose sur un processus « branché » conduisant au fil des générations cellulaires à une forte diversité clonale et donc à une hétérogénéité intratumorale [25]. Les données récentes de séquençage sur tumeur ou sur cellules uniques soutiennent cette hypothèse dans une large mesure [26].

La comparaison des profils mutationnels des différents sous-clones d'une tumeur primaire et de ses lésions métastatiques a également permis de démontrer l'hétérogénéité spatiale au sein d'une lésion et de déduire l'ordre dans lequel les événements oncogéniques se sont produits [27]. Ces travaux ont notamment conduit à la mise en évidence d'anomalies restreintes aux lésions métastatiques et à la caractérisation d'anomalies présentes dans certaines régions de la tumeur primaire et absentes dans les métastases étudiées, démontrant une évolution parallèle et suggérant une dissémination précoce de certaines cellules cancéreuses.

L'hétérogénéité génétique spatiale intratumorale a des implications considérables pour le diagnostic et le pronostic, les analyses d'une biopsie unique pouvant conduire à sous-estimer la complexité du paysage de mutations somatiques d'une tumeur et de ses dérivés métastatiques. Par ailleurs, l'hétérogénéité génétique et la plasticité phénotypique des cellules cancéreuses représentent des défis majeurs pour la conception de thérapeutiques ciblées. En théorie, la cible idéale est représentée par un événement oncogénique causal présent dans toutes les cellules tumorales. Cependant, si la croissance tumorale repose sur plusieurs événements ramifiés, la résistance aux médicaments anticancéreux peut survenir au travers de la sélection d'un sous-clone rare préexistant [28,29]. Une approche systématique de l'hétérogénéité moléculaire dans le cancer est aujourd'hui nécessaire pour évaluer son réel impact sur la prise en charge thérapeutique [30]. Les analyses longitudinales, et exhaustives, d'une lésion primaire et de ces éventuelles lésions métastatiques étant complexes à mettre en place et onéreuses, de nouvelles approches méthodologiques ont été développées, en particulier la détection par séquençage de l'ADN tumoral libre plasmatique circulant et de l'ADN de cellules tumorales circulantes [31,32].

Big Data et recherche clinique en cancérologie

Initialement restreints aux programmes de recherche, les progrès technologiques et la diminution des coûts du séquençage ont permis d'envisager la production de volumes importants de données en milieu clinique, chez des patients en situation de rechute et résistants aux traitements classiques, promesse d'une médecine personnalisée basée sur une analyse des caractéristiques de la tumeur de chaque patient et de la constitution génétique du patient lui-même. Les séquenceurs utilisés dans le domaine de la cancérologie sont principalement issus des technologies Illumina ou Ion Torrent. Dans les deux cas, ils génèrent des lectures courtes (100 pb environ) et offrent une bonne sensibilité de détection des mutations. De nouvelles technologies comme celles proposées par Oxford Nanopore Technologies (*e.g.* MinION) produisent à faible coût également des lectures de plusieurs dizaines de kilobases et sont actuellement en développement actif. Même si elles souffrent encore d'un taux d'erreur important pour les substitutions et les indels (12 %) [33], elles semblent adaptées pour la recherche de grands réarrangements comme, par exemple, les fusions géniques. Ces technologies ne supplanteront donc pas celles en place en recherche clinique mais pourront éventuellement être utilisées en complément. Actuellement, la grande majorité des études s'intéresse à un nombre restreint de gènes (1 à 500), mais récemment, des données d'exome entier ou de transcriptome ont pu être produites dans le cadre d'études cliniques [34]. La proposition, en juin 2016, du plan « France Médecine Génomique 2025 » (http://www.gouvernement.fr/sites/default/files/document/document/2016/06/22.06.2016_remise_du_rapport_dyves_levy_-_france_medecine_genomique_2025.pdf) visant à instaurer la généralisation de ces approches (WGS, WXS et RNASeq) dans le cadre d'un « parcours de soins générique avec un accès commun à tous les patients affectés par les cancers, maladies rares ou communes » va également résolument dans cette direction et contribue à effacer les limites entre les cadres recherche et clinique dans la production et

l'exploitation des NGS. Ce plan vise, par exemple, à la prise en charge à l'horizon 2020 de près de 235 000 séquences de génomes par an, soit une production annuelle de plusieurs dizaines de Péta octets, plus d'un ordre de grandeur au-dessus de ce qu'ont fourni les programmes de recherche du type TCGA ou ICGC. Ceci représente bien évidemment, en plus d'un enjeu de Santé publique, un formidable défi organisationnel.

Pour l'heure, la très grande majorité des études cliniques inclut un nombre plus restreint de patients, généralement en situation métastatique ayant échappé aux traitements standards. On distingue les essais paniers (*basket studies*) qui permettent un screening moléculaire ciblé sur l'altération génomique d'un gène ; on peut citer par exemple le programme national AcSé vemurafenib qui permet d'inclure des patients ayant des pathologies diverses mais en commun une altération génomique de BRAF. Les études de type « screening moléculaire haut débit » (*umbrella studies*) consistent en une caractérisation d'un panel de gènes dont les résultats permettent dans 30 à 40 % des cas d'identifier une altération « actionnable », c'est-à-dire définie comme étant la cible d'une molécule ou d'une combinaison de molécules disponibles dans le cadre d'essais cliniques de phase I-II. On peut citer à titre d'exemples les programmes MOSCATO (*MOlecular Screening for CAncer Treatment Optimization*) ou ProfiLER (Profilage LYric Et Région). Enfin, les essais dits comparatifs (*proof of concept*) comparent l'intérêt des approches expérimentales basées sur le profilage génomique des tumeurs des patients à des approches classiques. On peut citer par exemple les essais SHIVA [35], SAFIR02 et MOST (*My Own Specific Therapy*).

Ces programmes ont été à l'origine d'une réorganisation des établissements spécialisés (*comprehensive cancer centers*) avec, d'une part, la mise en place de réunions multidisciplinaires dites « moléculaires » (*molecular tumor board*) qui s'ajoutent à celles classiquement organisées par organe et incluant des compétences nouvelles (bio-informaticiens, biologistes). Elle a aussi nécessité un investissement conséquent dans des équipements permettant la production des données et une collaboration étroite avec des plateformes de bio-informatique pour le traitement et le stockage des données, autant d'éléments nécessitant une adaptation de l'ensemble des intervenants.

Le nécessaire partage des données générées dans le cadre de ces essais cliniques de médecine personnalisée est l'objet d'initiatives internationales et nationales [36,37]. Elles doivent permettre d'identifier des associations entre des altérations rares et la réponse à des thérapies ciblées, qui ne pourra se faire que si la communauté médico-scientifique décide de mettre en commun leurs données. De nombreux obstacles techniques, éthiques, méthodologiques et réglementaires existent. En France, un effort national sous l'égide de l'INCa et baptisé OSIRIS incluant les huit SIRICs, s'est donné pour objectif le partage des données cliniques et génomiques produites au plan national par différents programmes de médecine personnalisée (∼4000 patients). Toutes ces actions ont été rendues possibles grâce à un investissement fort de l'ensemble des acteurs, médecins, chercheurs, informaticiens, bio-informaticiens, ingénieurs, biologistes, qui interagissent ensemble hors de leurs schémas habituels. Enfin, il ne faut pas oublier que l'espoir généré par ces Big Data en cancérologie n'est permis que par le consentement et la participation active des patients à ces essais de screening moléculaire, rappelant la nécessité d'information et de sensibilisation du grand public sur leurs intérêts et leurs enjeux.

Conclusions et perspectives

Si les perspectives des analyses moléculaires, et des NGS en particulier, en cancérologie suscitent un intérêt considérable tant en termes de recherche qu'en termes de prise en charge des patients, elles représentent également des défis importants pour la communauté biomédicale. Le premier défi concerne le risque de considérer le tissu cancéreux comme un milieu stable et homogène au plan génétique et épigénétique, en omettant de prendre en compte la plasticité des cellules cancéreuses et leur diversité, sources majeures de complexité, d'échecs thérapeutiques et de rechutes. Il est clair aujourd'hui que si l'analyse moléculaire d'une tumeur constitue une étape essentielle de sa caractérisation biologique, elle ne peut à elle seule permettre de prédire son évolution clinique. Aussi, la vision réductrice du « tout génétique » pourrait, si l'on n'y prend garde, devenir un frein pour les analyses fonctionnelles et les études d'interaction entre cellules cancéreuses et cellules de l'environnement tumoral, indispensables à une réelle compréhension biologique de la tumeur. De ce point de vue, le développement de modèles mathématiques dynamiques, permettant de décrire l'évolution temporelle de la tumeur et munis de paramètres « personnalisés » par une caractérisation moléculaire propre à chaque patient, constituera une clef indispensable. Un second défi concerne le caractère éminemment pluridisciplinaire de ces activités, que ce soit au plan recherche ou clinique. L'instauration du dialogue et l'orchestration des échanges entre les différents acteurs : médecins, biologistes, bio-informaticiens, biomathématiciens constituent sans aucun doute une étape déterminante et nécessitent de revoir en profondeur nos formations universitaires et médicales, nos pratiques et, au fond, nos modes même de pensée.

Remerciements : LYric Grant INCa-DGOS-4664.

Références

[1] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets – update. Nucleic Acids Res 2013;41(Database issue):D991–5.

[2] Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update – simplifying data submissions. Nucleic Acids Res 2015;43(Database issue): D1113–16.

[3] Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature 2013;502(7471):333–9.

[4] Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International Cancer Genome Consortium Data Portal – a one-stop shop for cancer genomics data. Database (Oxford) 2011;2011:bar026.

[5] Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature 2013;500(7463):415–21.

[6] Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 2016; 534(7605):47–54.

[7] Ferrari A, et al. A whole genome sequence and transcriptome perspective on HER2-positive breast cancers. Nat Commun 2016;7:12222. http://dx.doi.org/10.1038/ncomms12222.

[8] Jiang P, Liu XS. Big data mining yields novel insights on cancer. Nat Genet 2015;47(2): 103–4.

[9] Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012;2(5):401–4.

[10] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6 (269):pl1.

[11] Lee H, Palm J, Grimes SM, Ji HP. The Cancer Genome Atlas Clinical Explorer: a web and mobile interface for identifying clinical-genomic driver associations. Genome Med 2015;7:112.

[12] Wan YW, Allen GI, Liu Z. TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. Bioinformatics 2016;32 (6):952–4.

[13] Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature 2012;490(7418):61–70.

[14] Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486(7403):346–52.

[15] Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. Nat Med 2015;21(11):1350–6.

[16] Cancer Genome Atlas Research N. The molecular taxonomy of primary prostate cancer. Cell 2015;163(4):1011–25.

[17] Bunn Jr PA, Franklin W, Doebele RC. The evolution of tumor classification: a role for genomics? Cancer Cell 2013;24(6):693–4.

[18] Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014;511(7511):543–50.

[19] Clinical Lung Cancer Genome P, Network Genomic M. A genomics-based classification of human lung tumors. Sci Transl Med 2013;5 (209):209ra153.

[20] Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012;489 (7417):519–25.

[21] Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45(10):1113–20.

[22] Iglesia MD, Parker JS, Hoadley KA, Serody JS, Perou CM, Vincent BG. Genomic analysis of immune cell infiltrates across 11 tumor types. J Natl Cancer Inst 2016;108(11).

[23] Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12(5):453–7.

[24] Nowell PC. The clonal evolution of tumor cell populations. Science 1976;194(4260): 23–8.

[25] Greaves M, Maley CC. Clonal evolution in cancer. Nature 2012;481(7381):306–13.

[26] Swanton C. Intratumor heterogeneity: evolution through space and time. Cancer Res 2012;72(19):4875–82.

[27] Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 2012;366(10):883–92.

[28] Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature 2012;481(7382):506–10.

[29] Roche-Lestienne C, Lai JL, Darre S, Facon T, Preudhomme C. A mutation conferring resistance to imatinib at the time of diagnosis of chronic myelogenous leukemia. N Engl J Med 2003;348(22):2265–6.

[30] Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. Br J Cancer 2013;108(3):479–85.

[31] Heitzer E, Auer M, Gasch C, Pichler M, Ulz P, Hoffmann EM, et al. Complex tumor genomes inferred from single circulating tumor cells by array-CGH and next-generation sequencing. Cancer Res 2013;73(10): 2965–75.

[32] Lo YM, Chiu RW. Plasma nucleic acid analysis by massively parallel sequencing: pathological insights and diagnostic implications. J Pathol 2011;225(3):318–23.

[33] Ip CL, Loose M, Tyson JR, de Cesare M, Brown BL, Jain M, et al. MinION analysis and reference consortium: phase 1 data release and analysis. F1000Res 2015;4:1075.

[34] Blay JY, Tredan O, Ray-Coquard I, Rivoire M, Mehlen P, Puisieux A, et al. [Quinze questions importantes a se poser en oncologie en 2015]. Bull Cancer 2015;102(6 Suppl. 1): S22–6.

[35] Le Tourneau C, Delord JP, Goncalves A, Gavoille C, Dubot C, Isambert N, et al. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. Lancet Oncol 2015;16(13):1324–34.

[36] Kuehn BM. Alliance aims for standardized, shareable genomic data. JAMA 2013;310 (3):248–9.

[37] Rose S. Huge data-sharing project launched. Cancer Discov 2016;6(1):4–5.